Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 182 – 187

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

Spoken Language Identification with Phonotactics Methods on Minangkabau, Sundanese, and Javanese Languages

Nur Endah Safitri, Amalia Zahra, and Mirna Adriani*

Universitas Indonesia, Depok, 16424, Indonesia

Abstract

Research in the field of spoken language identification (spoken LID) on local languages helps to extend the outreach of technology to local language speakers. This research also contributes to the preservation of local languages. In this paper, we report our work on identifying spoken data in three local Indonesian languages: Minangkabau, Sundanese and Javanese. Statistical phonotactics models are created to map the speech signals into the language used by the speaker. We use two phonotactics methods, namely Phone Recognition followed by Language Modelling (PRLM) and Parallel Phone Recognition followed by Language Modelling (PPRLM). PRLM method shows the highest accuracy using the phone recognizer trained for English and Russian with the average of 77.42% and 75.94% respectively.

Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords: spoken language identification; phonotactic methods

1. Introduction

Spoken Language Identification is the process of matching a stream of spoken sound waves to the language they are using. There are a number of methods which can be implemented, such as acoustic based methods and phonotactics based methods. This research tested the accuracy of phonotactics modelling for language identification on Minangkabau, Sundanese and Javanese languages.

The number of local language speakers in Indonesia is considered high, which is explained by the large number of ethnic groups present in Indonesia. The Central Statistical Bureau (Badan Pusat Statistik - BPS), the government-run

* Corresponding author.
E-mail address: mirna@cs.ui.ac.id

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.047

Indonesian statistics bureau, reported that there are 1.340 ethnic groups in Indonesia as of 2010. Furthermore, almost 80 % of people above the age of 5 use at least one local language on daily basis[1]. Indonesia has the second largest number of living languages in the world, which constitutes almost 10% of the world languages[2]. These numbers shows that there is great potential in expanding the reach of technology, with its attendant benefits, if local languages can be used as a technological interface. Voice interface technology is crucial for people who are cannot use text based interface due to blindness or illiteracy. According to BPS, there are more than 5% or about 13 million people in Indonesia who are illiterate[3]. The utilization of voice interface will make it easier for those 13 million people to benefit from technology. Unfortunately, the amount of local language speakers is in decline. According to the BPS report in 2010, the number of people using the national language as first language doubled in the past 20 years, illustrating that the usage of local language is decreasing. UNESCO reported that there are 146 Indonesian local language that have become extinct or endangered between 1950 to 2010[4]. Therefore, it is necessary to take actions to preserve local languages.

There has been some research in the field of spoken language identification by phonotactics. In 2006, Pavel Matejka, et al. have used Phone Recognition followed by Language Modelling (PRLM) and Parallel Phone Recognition followed by Language Modelling (PPRLM) to differentiate twelve languages by using trigrams and likelihood estimation[5]. They found that PRLM systems with more training data outperform PPRLM systems. The result also pointed out that it is better to use less phone tokenizer with more accuracy than to use more phone tokenizer with lower accuracy.

This paper explains the process of building a spoken LID system for Indonesian local languages. We apply the PRLM and PPRLM approaches since those approaches has been successfully used in other languages. This research also covers experiments on how to effectively use universal phone recognizer to implement phonotactics methods, in order to create a spoken LID system. This research utilized independently collected speech corpora, consisting of three of the Indonesian local languages: Minangkabau, Sundanese and Javanese.

Following this section are Section 2, 3, 4, 5, and 6. Section 2 describes the phonotactics methods which are implemented. Section 3 explains the phone recognizer and statistical model builder which are utilized in this research. Section 4 reports on how the speech corpora are gathered and prepared. Section 5 explains the experiments that have been done along with their results and analysis. The last section contains the conclusion based on analysis of the experiments.

2. Phonotactic Methods for Spoken Language Identification

Spoken LID is part of signal processing on a stream of speech signal. Using the target languages as the classes, spoken LID can also be considered as a classification problem. The features that can be used to build a classifier model could be a phrase, word, syllable, phoneme, phone, or the frequency variation of the speech signal itself. Spoken LID techniques can be differentiated by the complexity of the features used to build the classifier model.
From the lowest to highest level of abstraction, there are acoustic, prosodic, phonotactics, lexical, and syntactic methods[6]. This research uses and compares the performance of two phonotactics methods: PRLM and PPRLM.

2.1. Phone Recognition Followed by Language Modelling

Spoken LID system with Phone Recognition followed by Language Modeling (PRLM)[7] approach first recognizes the phone from the stream of speech signal, and then classifies the phone to the target languages using statistical modeling. This system uses a single phone recognizer, regardless of what language the speech signal is spoken in. This phone recognizer is used as a universal recognizer, by creating an n-gram statistical model. This is done by calculating the likelihood of sequences of phones appearing in a certain language.

The phones from the speech signal is recognized, which allows log likelihood values to be calculated from the statistical model for each language. The log likelihood maximum value determines which language is used by the speaker.

2.2. Parallel Phone Recognition Followed by Language Modelling

The main difference between PRLM with PPRLM is in the number of phone recognizer used. In

PPRLM method[7], the phones from the speech signal is recognized by a few phone recognizers. These phone recognizers are trained to identify phones from different languages. The statistical model for each language is made from the result of interpolated models built by the phone from each phone recognizer.

PPRLM uses multiple phone recognizer. Streams of phone symbols from a speech signal are recognized by each of the phone recognizer. Then, log likelihood values for each stream of phone symbols are calculated from each interpolated *n*-gram language models. Log likelihood values calculated from a language model for each stream of phone symbols are combined and compared with the combined log likelihood values from all other language models. The maximum value determines which language is used by the speaker.

3. Phone Recognizer and Statistical Model Builder

The phone recognizer used in this research is PhnRec, a system developed by Brno University of Technology. PhnRec has been used as phone recognizer in a spoken LID system for six languages[5]. Those language are English, Germany, Hindi, Japanese, Mandarin, and Spanish. PhnRec has been developed to recognize the phones in four languages. They are Czech, English, Hungarian and Spanish. PhnRec is trained with TIMIT and SpeechDat-E speech corpora. PhnRec is using HTK Label File or Master Label File format for output. We use the PhnRec for our work since we have had a corpus for building phone recognizer in our local languages.

After the streams of phones are recognized, an n-gram statistical model is built with SRILM. SRILM is a statistical language modeling toolkit containing executable files and C++ library. SRILM is developed by Speech Technology and Research Laboratory of SRI International. The statistical model built with SRILM uses the ARPA format[8]. SRILM toolkit is also used to calculate log likelihood values from the statistical model.characters

4. Preparation of Speech Corpora

The speech corpora consist of three Indonesian local languages: Minangkabau, Sundanese and Javanese. The three languages are considered as the mostly spoken languages in Indonesia. Minangkabau is spoken in West Sumatra island, Sundanese is spoken in West Java island, and Javanese is spoken in Central and East Javanese island. The three languages have their own characters. There are six speakers for each language, 18 speakers in total. The speakers are native, being proficient as well as coming from the area where the language originates. Each language has three male and three female speakers. The gender is balanced to ensure that male or female-specific speech features do not create any erroneous correlations in the LID calculations. The speech signal is recorded into single channel audio files with 16 bit sample rate and 16 kHz sample size.

Table 1 Speech Corpora Duration

Local language	#script words	Average duration of speaker	Total speech corpus duration
Minangkabau	4006	42 minutes	252 minutes
Sundanese	4356	39 minutes	236 minutes
Javanese	4196	41 minutes	246 minutes

The speech corpus is a recording of the speakers reading a script in local languages. The scripts consists of articles or short stories in local languages. The duration of each speech corpus is listed in Table 1. The speech corpora is divided into three sets for experimentation: Training, Development, and Test. The proportion of this partition is explained in the Table 2.

Table 2 Set Partition for Experiments

Set name	# of speaker	Gender division	Average Set Duration per Language
Training Set	3	2 male 1 female or 2 female 1 male	124 minutes
Development Set	1	1 female or 1 male	42 minutes
Test Set	2	1 female and 1 male	79 minutes

In addition, we also calculate word frequency in corpus written in Javanese. This value will also considered in choosing the right possible segmented word. Table 3 shows five words and its frequency.

5. Experiments and Discussion

The experiment is done on each n-gram statistical model for PRLM and PPRLM methods. The accuracy of those models is calculated with confusion matrices. The values which are compared are accuracy, false positive rate, precision, and recall. The variable which are altered in the PRLM and PPRLM method is the n value for n-gram statistical model, with n=3 to n=10 inclusive.

5.1 PRLM Experiments

In PRLM experiments, systems are differentiated by which phone recognizer is used. Four spoken language identification systems are tested, with each being trained in Czech, English, Hungarian and Spanish. Each system is tested with 3-gram to 10-gram statistical models.

5.2 PPRLM Experiments

PPRLM experiments involve two spoken language identification systems. The first system uses all of the phone recognizer available in PhnRec—Czech, English, Hungarian and Spanish—in order to create interpolated models and tokenize phones. The second system uses two of the phone recognizers which yield the best accuracy in PRLM experimentation. English and Russian phone recognizers are selected according to their performance in PRLM systems. The accuracy results of these systems are also compared and tested with 3-gram to 10-gram statistical models.

Table 3, 4, 5, and 6 show the comparison of PRLM and PPRLM.

Table 3 Accuracy Comparison of PRLM and PPRLM methods for Spoken LID

n-gram	Methods					
	PRLM with phone recognizer trained on:				PPRLM with phone recognizer trained on:	
	CZ	EN	HU	RU	All	EN-RU
3-gram	58.06	74.19	62.37	72.04	74.19	70.97
4-gram	56.99	76.34	54.84	79.57	73.12	66.67
5-gram	59.14	76.34	56.99	80.65	73.12	66.67
6-gram	59.14	76.34	55.91	77.42	75.27	67.74
7-gram	59.14	75.27	55.91	77.42	74.19	68.82
8-gram	60.22	76.34	54.84	77.42	74.75	67.11
9-gram	59.14	76.34	53.76	77.42	73.56	71.43
10-gram	56.99	76.34	54.84	77.42	73.12	70.97
$\bar{x}$ *Accuracy for all n*	58.60	75.94	56.18	77.42	73.92	68.80

As shown in Table 3, the accuracy results of PRLM systems with English or Russian phone recognizers are higher than the ones with Czech or Hungarian phone recognizers. Furthermore, PPRLM systems with selected phone recognizers turned on (EN-RU) turn out to be less accurate than PPRLM systems with all phone recognizers active (CZ-EN-HU-RU).

Table 4 False Positive Rate Comparison of PRLM and PPRLM methods for Spoken LID

n-gram	Methods					
	PRLM with phone recognizer trained on:				PPRLM with phone recognizer trained on:	
	CZ	EN	HU	RU	All	EN-RU
3-gram	30.53	19.32	28.18	21.14	19.77	21.44
4-gram	32.05	17.80	33.79	15.76	20.61	24.92
5-gram	30.53	17.80	32.20	14.92	20.61	24.92
6-gram	30.68	17.95	32.88	17.27	18.94	24.09
7-gram	30.53	18.71	32.88	17.27	19.77	23.26
8-gram	29.70	17.95	33.64	17.27	19.74	23.17
9-gram	30.45	17.95	34.39	17.27	19.87	22.99
10-gram	31.97	17.95	33.56	17.27	20.53	21.67
$\bar{x}$ *False Positive rate for all n*	30.80	18.18	32.69	17.27	19.98	23.31

Table 4 shows that the false positive rate of PRLM systems with English and Russian phone recognizers active are better than the ones with Czech and Hungarian phone recognizers. In line with the accuracy results in Table 3, PPRLM systems with selected phone recognizers (EN-RU) return more false positive results than PPRLM systems with all phone recognizers (CZ-EN-HU-RU).

Table 5 Precision Comparison of PRLM and PPRLM methods for Spoken LID

n-gram	Methods					
	PRLM with phone recognizer trained on:				PPRLM with phone recognizer trained on:	
	CZ	EN	HU	RU	All	EN-RU
3-gram	45.36	57.94	45.31	64.68	75.39	60.14
4-gram	35.28	59.12	29.96	75.26	62.17	59.35
5-gram	38.11	61.02	33.83	77.24	61.74	59.35
6-gram	38.18	57.70	32.76	69.86	63.77	61.11
7-gram	38.39	56.27	32.76	69.86	62.69	61.59
8-gram	40.30	57.86	31.18	69.86	63.30	59.31
9-gram	38.97	57.86	29.43	69.86	61.92	67.65
10-gram	36.54	57.70	31.32	69.86	61.44	63.94
$\bar{x}$ *Precision for all n*	38.89	58.19	33.32	70.81	64.05	61.56

Table 5 shows that the best precision is produced by PRLM systems with Russian phone recognizers. Both PPRLM systems yield better results compared with the other PRLM systems.

Table 6 Recall Comparison of PRLM and PPRLM methods for Spoken LID

n-gram	Methods					
	PRLM with phone recognizer trained on:				PPRLM with phone recognizer trained on:	
	CZ	EN	HU	RU	All	EN-RU
3-gram	37.21	59.26	42.93	57.24	60.61	55.72
4-gram	35.02	61.95	31.31	67.17	57.74	48.65
5-gram	37.71	61.95	34.34	69.02	57.74	48.65
6-gram	37.71	61.28	32.83	64.48	60.77	50.51
7-gram	37.71	59.76	32.83	64.48	59.26	52.02
8-gram	39.56	61.28	31.31	64.48	58.77	50.26
9-gram	38.05	61.28	29.80	64.48	59.72	54.20
10-gram	35.02	61.28	31.65	64.48	58.08	55.39
$\bar{x}$ *Recall for all n*	37.25	61.01	33.38	64.48	59.09	51.93

Table 6 shows that the recall comparison of each system show the same pattern in both accuracy and false positive rate. The best recall results are returned by PRLM systems with Russian and English phone recognizer. However, the average recall for all PPRLM systems are still better than the average of all PRLM systems.

The accuracy, false positive, precision and recall results above illustrate that the change of value n for n-gram statistical model does not impart any meaningful change in accuracy. In sing the correct phone recognizer is crucial in building a high accuracy system for Spoken LID with phonotactics methods. The general trend is that PPRLM systems still yield better results than PRLM.

The experiment uses corpora that are self-gathered, so the noise level is still too high in some parts of the corpus, which may have influenced the result of the experiments.

6. Conclusions

Phonotactics methods can indeed be used to create Spoken LID systems for Minangkabau, Sundanese and Javanese languages. The test results indicate that the variation for n in n-gram statistical model from 3 to 10 does not cause any significant change in accuracy. Nonetheless, experiments show that the best value for n in n-gram statistical modelling is 3 or 5.

Accuracy of Spoken LID system with PRLM and PPRLM methods is affected more by the performance of phone recognizer that is used. The highest accuracy with PRLM method is obtained when phone recognizers trained for English and Russian language are used, with the average accuracy of 77.42% and 75.94% respectively. However, in this experiment, PPRLM average accuracy is 73.92% while the PRLM method average is 67.04%, which clearly shows that PPRLM is the more accurate method of the two.

There is much room for improvement, as smoothing parameters have not yet been applied to the statistical model, and the speech corpora can be expanded to incorporate speakers of various ages in order to ensure that age-specific spoken signal features do not skew the LID estimation.

References

1 Sub Direktorat Statistik Politik dan Keamanan, "Statistik Politik 2014," Badan Pusat Statistik, Jakarta, 2014.

2 Lewis, M. Paul, G. F. Simons and C. D. Fennig, Eds., "Ethnologue: Languages of the World, Eighteenth edition," Dallas, Texas: SIL International, 2015.

3 Badan Pusat Statistik, "Persentase Penduduk Berumur 10 tahun Ke atas yang Buta Huruf menurut Provinsi dan Jenis Kelamin, 2009-2013," 2015.

4 C. Moseley, Ed., *Atlas of the World's Languages in Danger,* 3 ed., Paris: UNESCO Publishing, 2010.

5 P. Matejka, P. Schwarz, J. Cernocky and P. Chytil, "Phonotactic Language Identification using High Quality Phoneme Recognition," *Eurospeech 2005,* September 2005.

6 R. Tong, B. Ma, D. Zhu, H. Li and E. S. Chng, "Integrating Acoustic, Prosodic and Phonotactic Features for Spoken Language Identification," *International Conference on Acoustic, Speech, and Signal Processing,* pp. 205-208, 2006.

7 M. A. Zissman, "Comparison of Four Approach to Automatic Language Identification of Telephone Speech," *IEEE Transactions on Speech and Audio Processing,* vol. 4, no. 1, pp. 31-44, January 1996.

8 A. Stolcke, "ngram-format Man page," 2004. [Online]. Available: http://www.speech.sri.com/projects/srilm/manpages/ngram-format.5.html. [Accessed 2015].